Robots as Ideal Moral Agents per the Moral Responsibility System

Gogoshin, Dane Leigh

IOS PRESS
2020-12

Gogoshin , D L 2020 , Robots as Ideal Moral Agents per the Moral Responsibility System . in M Nørskov , J Seibt & O S Quick (eds) , Culturally Sustainable Social Robotics : Proceedings of Robophilosophy 2020 / TRANSOR 2020 . Frontiers in Artificial Intelligence and Applications , vol. 335 , IOS PRESS , Amsterdam , pp. 525-534 , International Research Conference Robophilosophy 2020 , Aarhus , Denmark , 18/08/2020 . https://doi.org/10.3233/FAIA200952

http://hdl.handle.net/10138/325342
https://doi.org/10.3233/FAIA200952

acceptedVersion

Robots as Ideal Moral Agents per the Moral Responsibility System

Dane Leigh GOGOSHIN [a,1]
[a] *University of Helsinki*

Abstract. Contrary to the prevailing view that robots cannot be full-blown members of the larger human moral community, I argue not only that they can but that they would be ideal moral agents in the way that currently counts. While it is true that robots fail to meet a number of criteria which some human agents meet or which all human agents could in theory meet, they earn a perfect score as far as the behavioristic conception of moral agency at work in our moral responsibility practices goes.

Keywords. Moral responsibility, ethical behaviorism, robot moral agency

1. Introduction

According to John Danaher's recent ethical behaviorism thesis [1], a robot should be afforded moral status if it is *roughly performatively equivalent* to another agent on whom we confer moral status. A robot will have the *rough performative status* of a sentient being on whom we confer moral status for its capacity to feel pain, for example, if the robot behaves in a way that we recognize as an expression of pain. Such "a robot should be granted the same moral status as any other entity to whom we ascribe moral status on the grounds that they can feel pain" [1:4]. Mark Coeckelbergh has argued for a similar approach to moral agency and responsibility, *virtual* agency and responsibility, in which he proposes to "redirect our attention to the various ways in which non-humans, and in particular robots, appear to us as agents, and how they influence us in virtue of this appearance" [2:181] and away from the traditional approach which grounds those concepts on agents' mental content. In 2004, Luciano Floridi and J. W. Sanders [3] put forth the concept of a "mind-less morality" in which moral agency is conferred on the basis of "observables" and not on "psychological speculation" [3:365] (a view upon which John Sullins builds in [4]).

Appealing as this shift away from the opacity of the inner human and the difficulty of concepts connected to moral responsibility (e.g. free will, consciousness, intentionality, responsibility) may be, intuitively "what's going on on the inside" seems to matter a lot and perhaps it should. On one hand, in certain contexts and within certain of our moral communities, it does, at least subjectively. On the other hand, for society at large – the most universal of our moral communities – I argue that whether or not it should, it does not and that this is consistent with our moral responsibility practices. What

1 Corresponding Author: Dane Leigh Gogoshin, Department of Practical Philosophy, University of Helsinki, P.O. Box 54, 00014 Helsinki, Finland; e-mail: dane.gogoshin@helsinki.fi.

matters for this level of moral community membership is our "moral performance" – that we behave in conformity with society's rules and values.

This paper investigates the standards for moral agency underlying our moral responsibility practices and determines that robots could theoretically attain them, provided one accepts the minimal conception of morality presented below and that robots can be programmed accordingly,[2] for it turns out that those standards are much narrower than those we tend to associate with moral agency. At stake in the machine ethics literature (see e.g. the one at issue in [9]) is a more demanding conception of moral agency than the one behind our moral responsibility system. My thesis diverges from those of both Danaher and Coeckelbergh in that (a) it does not prescribe ethical behaviorism or virtual agency/responsibility; it acknowledges that the conception behind them underlies the moral responsibility system. (b) It locates the relevant moral performance to rule/value compliance. The basis for (b) is to be found in the minimal, universal requirements for community, for membership into the larger society's "moral agents club," not necessarily for certain specialized settings or intimate relationships. (c) As far as the rights and duties this membership entails, it assumes minimally only that it grants access to general public settings and renders members subject to expectations for moral behavior in those settings.

In the first and lesser known of his two papers dedicated to moral philosophy, "Social Morality and the Individual Ideal" [10:5], P. F. Strawson presents a minimal conception of morality as follows:

> Now it is a condition of the existence of any social organization, any human community, that certain expectations on the part of its members should be pretty regularly fulfilled; that some duties, one might say, should be performed, some obligations acknowledged, some rules observed. We might begin by locating the sphere of morality here. It is the sphere of observation of rules, such that the observance of some such set of rules is the condition of the existence of society. This is a minimal interpretation of morality. It represents it as what might literally be called a kind of public convenience: of first importance as a condition of everything that matters, but only as a condition of everything that matters, not as something that matters in itself.

According to Strawson, then, morality in its most basic terms[3] – the observance of a certain set of rules which make society possible – makes possible the higher human goods. A moral agent is thus, first and foremost, an agent who follows and whom we expect to follow these rules.[4] Why agents comply with the rules is of, at most, secondary concern.[5] That's not to say that society does not care about or seek to influence our mental content but it does so with the goal of influencing our behavior. Mental content is relevant to our behavior – it determines how an agent complies with rules and values. But it is possible to comply with them simply by, e.g., conforming to our peers' behavior, in order to avoid negative consequences, and/or to obtain rewards. Why and how we

2 See [5] for a proposal of a rule-based (Kantian) machine. See [6] and [7] for an optimistic view of programmable ethics and [8] for a pessimistic one.

3 He acknowledges the inadequacy of this minimal conception of morality, but sees "considerable merit" in it as well [10:5].

4 The negative reactive attitudes (Strawson [11:9-12]), in first terms, are reactions to the disappointment of the expectations to which we hold all non-exempt (highly abnormal or immature) agents (cf. [12:11]).

5 Only certain legal violations result in a legally sanctioned inquiry into our mental content as reasons and motives for action. We also tend in the inter-personal, even while caring about mental content, to solicit it (i.e. asking about intentions, beliefs, motives) only in cases of perceived wrongdoing.

comply with society's rules and respect its values are not important for membership into the larger moral community, only *that* we do.

My claim for the kind of moral agency at work in the moral responsibility system rests on the following four considerations. (i) The moral responsibility practices of blame, praise, punishment, reward and the reactive attitudes[6] target behavior which exceeds or transgresses society's rules and values, silently approving (in practice) behavior which merely conforms to those rules and values. An agent who may be theoretically blameworthy for doing the right thing (or avoiding wrongdoing) for other than the moral reason, tends not to be blamed in practice. (ii) These practices amount to behavioral conditioning practices. They appeal to our basic psychosocial sensitivities (sensitivities to our own and others' pain and pleasure, to our desire for approval, praise, and reward and to our distaste for disapproval, blame, and punishment), pairing positive behavior with positive responses and negative behavior with negative responses. In this way, they reinforce behavior which conforms with rules/values – irrespective of an agent's reasons for acting – and deter behavior which does not. (iii) Behavioral conditioning can address only a very limited set of behaviors. (iv) In order to legitimately hold wrongdoers to account for their actions, one needs a high degree of confidence in their guilt. Such confidence is only possible when the criteria are sufficiently low.

A final preliminary clarification is in order. On Strawson's view of morality as the observance of the rules that make community possible, rule-abiding robots meet the preconditions for "everything that matters" to humans (the "higher goods"). It may well be that "higher" levels of agency are required for these higher goods, of which robots may not be capable, but it does not necessarily follow that their very basic moral agency would obstruct our access to those higher goods or undermine the principles of human dignity and flourishing (as set forth in [13] and [14]). Whether such a species of moral agency should be spoken of in these terms is not at stake in this paper (until the conclusion); the moral responsibility system's conception of moral agency is.

In the remainder of this paper, I shall proceed as follows. In Section 2, I clarify the thesis and describe the nature and scope of our moral responsibility practices. In Section 3, I defend the claim that ethical behaviorism is the conception at work in the moral responsibility system and why performance (behavior) is what matters most to moral community membership. In Section 4, I address objections that one might have even should one accept the minimal account of morality as rule-observance and robots' capacity for it, and my account of the narrow moral agency at work in the moral responsibility practices. I conclude in Section 5.

2. Clarifying the Claims

Contrary to the prevailing view that robots cannot be full-blown members of the larger human moral community,[7] I argue not only that they can but that they would be ideal moral agents according to our moral responsibility standards. While it is true that robots fail to meet a number of criteria which some human agents meet or which all human

6 I refer here specifically to overt acts or expressions of blame, praise, etc., not to private, mental acts.

7 See [15] and [16] for the view of robots as in an "in-between position" – neither instruments nor moral subjects. See [17] and [8] for views against robotic moral agency.

agents could in theory meet,[8] they earn a perfect score as far as the conception of behavioristic moral agency at work in our moral responsibility system goes. Although an agent must be substantively responsible to deserve our moral responsibility responses (in terms of properly identifying with their motives, e.g. [21], or rational control, e.g. [22], [23]), I argue that blame, praise, punishment, reward, and the reactive attitudes target our most basic psychosocial sensitivities, sensitivities which all neurotypical, appropriately socialized human agents possess. An apt target for these responses in terms of moral (re)formation need only be sensitive in these ways and responsive to (i.e. (re)formable by) these practices. Furthermore, even if one's intention in blaming or praising, punishing or rewarding, is to (e.g.) dole out just deserts, by pairing positive responses (praise, approval, reward) with moral behavior and negative responses (blame, disapproval, punishment) with immoral behavior, we are engaging in operant conditioning practices (reinforcement and punishment), and the kind of moral agency at which they are geared is behavioral.

Moreover, these responses are triggered specifically by behavior which either transgresses or exceeds our rules and values and they largely ignore the mental content of behavior which conforms to them. At a societal level, only when agents act against society's rules/values do we bother about their intentions, motives, beliefs, underlying emotional difficulties, etc.[9] When heroic firefighters compromise their well-being for the sake of others, we do not ask them, "Why did you do that?" "Did you risk your life to impress your boss?" We tend, in practice, to take it for granted that morally praiseworthy acts are done for their intrinsic goodness. We tend not to ask our children why they have begun to follow the rules ("Did you stop hitting your sister just to make us happy?"); we are satisfied, when they do, that they do. The moral responsibility practices are largely silent on and thus approving of "mere good-doings"[10] – behavior which conforms to society's rules and values irrespective of agents' reasons (e.g. to avoid punishment, to please peers and authority figures, to keep one's job, to maintain freedom, etc). In line with Danaher's "ethical behaviorism," according to which what counts from the ethical perspective is how an agent performs, not "what's going on on the inside," the conception of agency at work in the moral responsibility system is narrow and behavioristic: a moral agent is one who conforms to society's rules and values.

Not only then do these practices foster mere conformity with rules/values (over acting for moral reasons), their regulative scope is limited to the domain of past actions. We cannot influence (i.e. positively or negatively reinforce) via the moral responsibility practices behaviors which have not occurred. Of course, we can let others know that we will react a certain way in case they behave a certain way and thereby possibly influence future action. But first, this may be a weaker form of influence than direct, emotional responses. Second and more importantly, the sphere of influence is still severely limited – limited to that which we can anticipate and articulate. This form of engagement, though more personal, is similar to the ways in which our society manages our environments, placing limits and negative incentives on certain actions. If B. F. Skinner [24], who insists that behavior is not distinct from character, were correct, then it is conceivable that we (neurotypical, socially-sensitive agents) can form an acceptable moral character via these practices in a reasonable period of time, provided sufficient exposure to a

8 Criteria such as appropriate histories per Alfred Mele's [18] history-sensitive account of autonomy and responsibility (see [19]), personality and subjectivity (see [20]), etc.

9 Moral responsibility theories are clearly concerned with the mental content, but I argue that the practices are largely not.

10 A term introduced to me in conversation and coined, at least in Finnish, by Antti Kauppinen.

sufficient (but finite) range of contexts and situations.[11] If Skinner were wrong, then these practices aren't aiming at forming character but rather at ensuring consistent behavioral results. Either way, our moral responsibility practices reflect a satisfaction with (or perhaps prioritization of) this kind of moral formation.

In practice and at the level of the larger society, given an awareness of society's rules and values and a sufficient sensitivity to acts of moral approval and condemnation, an agent is treated as morally responsible. Provided such an agent behaves according to those rules and values, whatever their motives, whatever their mental content, said agent is a moral agent. The basis for my opening claim should now stand in clear relief: robots who are equipped with all of their community's rules and values and the capacity to apply them when relevant, supposing of course that this is a sufficient basis for basic moral agency, will behave more morally than human agents who, despite our best efforts, routinely fail to do so, and will thus be ideal members of our larger moral community. Human agents possess varying degrees of sensitivity to our conditioning methods and suffer easily from over-sensitization or desensitization and so the process itself is very far from optimal or reliable. Moreover, what goes on on the human inside very often interferes with our moral performance. And although robots need to be responsive to the relevant rule/value, that responsiveness need not be composed of human stuff. What is crucial is *that*, not how or why, they conform to the relevant values and rules.

3. Why Performance Matters Most

Admittedly, this standard for moral agency is unsatisfying. In terms of ideal moral agency and grounding moral responsibility attributions (i.e. determining when praise and blame are theoretically justified), the mental content matters a great deal. Although conditioning practices may habituate us to right behavior and make it possible for reasons-responsiveness to arise,[12] we must also arrive at an understanding of what makes a given rule or value right and to seek to change those which are not. Autonomous agency (per [18:13]), though perhaps a standard to which we should not hold agents morally responsible (see [26]), requires a stringent hierarchy of history-sensitive values, desires, and beliefs (an evaluative basis), judgments formed on this evaluative basis, and the executing of intentions based upon those judgments. We should strive for this level of agency.

In his most influential paper, "Freedom and Resentment" [11], P. F. Strawson argues persuasively that our moral responsibility practices amount to much more than devices for social regulation, that our reactive attitudes (e.g. resentment, indignation, gratitude) are integral to the inter-personal, target others' quality of will, and are predicated upon a wider array of beliefs (i.e. justice, desert, freedom) than that of their social utility. It would be wrong, he asserts, "to forget that these practices, and their reception, the reactions to them, really are expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our natures, they express them" [11:15]. Far from denying their regulative effects, Strawson's emphasis still serves, I argue, to shift our attention away from the ways in

11 If Piaget [25] were right, then the kind of heteronomous moral agency at work in the moral responsibility system is already developed in preschoolers.

12 Per Aristotle in the Nicomachean Ethics, Book II

which these practices are regulative, how they affect moral formation, and what they say about the conception of moral agency on which they rely. And despite the conception of morality he presents in his earlier paper [10], his argument contra the moral responsibility consequentialist (the "optimist") [11] distracts from the role that our reactive attitudes play in establishing and upholding the rules/values that make community possible. Our moral responsibility practices – acts and expressions of moral condemnation and approval – whatever else they do, serve that end.

For the day-to-day functioning of the larger society, the most universal of our moral communities, what matters the most[13] is that people conform in action to the established rules. Not only is it impractical to manage agents' mental content in this context,[14] it would also be taken as an infringement on individual liberty to do so. Whereas modern liberal societies are comfortable with directly managing our moral agency via social policies aimed at reducing wrongdoing by minimizing opportunities for it (e.g. obvious security cameras, locked doors and barred windows, restrictions on weapons purchases, strategic outdoor lighting, etc.), they are less likely to endorse overtly managing our mental content.

Kenneth Himma [27:23] points to the "other minds" problem – the difficulty of justifying consciousness in anyone other than self – which, as Coeckelbergh [2:182] also emphasizes, makes moral responsibility attributions on the basis of mental states all the more intractable. Coupled with the complexity of the mechanisms guiding human moral behavior, this problem motivates an alternative approach, one which Coeckelbergh argues we already employ with one another and our pets – virtual agency and virtual responsibility [2:184). These ascriptions are made "on the basis of how the other is experienced and appears to them," not on how they really are. As Danaher [1:5] points out, ethical behaviorism has an advantage over other ethical views in that it respects our epistemic limits. Although an ethical behaviorist might well accept that there are inner mental states which provide the ultimate metaphysical ground for our ethical principles, they claim that we have no means of knowing those states directly; we can know them only by way of their behavioral representations. "Behaviour is then, for practical purposes, the only insight we have into the metaphysical grounding for moral status" [1:6].

Per our moral and legal responses, it is only when an agent transgresses the relevant rule or value that society is tasked with the discovery and assessment of their intentions and motives. It is in society's interest that people first and foremost behave according to the rules – not that they do so because the rules are right.[15] Via an appeal to our shared psychosocial sensitivities, the moral responsibility practices, despite all the other things that they do and reflect and further, despite the fact that we may in our smaller moral communities (e.g. our families, our religious communities, our partnerships, etc.) solicit and attend to each other's mental content, serve to condition human behavior. We thereby drive moral formation in a certain direction – behavioral conformism. In fact, I argue that our moral responsibility practices not only do not rely on substantive agency (in the form of moral reasons-responsiveness, freedom, control, or rationality), they in some

13 In the current state of human development and per the moral responsibility and legal systems.

14 Impractical in the sense of resource-intensive and difficult-to-impossible to reliably ascertain.

15 Even if it is not in society's long-term "big picture" interest or in the interest of human flourishing/species survival to prioritize performance. For this, we presumably need "substantive agency" (something beyond behavioral conformism, involving, *inter alia*, moral reasons-responsiveness).

ways oppose its development. As a result of the conditioning process, we face the forks in the paths of the garden of life from very loaded perspectives. Successfully conditioned agents perceive the path that respects the rules as green and promising and the path which doesn't as littered with red flags and otherwise barren. Can such an unequal choice truly count as a choice at all? A robot will follow what we've established as the green path every single time – no matter what. Would we not want such community members?

Perhaps we wouldn't want such agents within our intimate communities. Perhaps robots should be treated like strangers in this respect; we should keep a healthy distance. But the strongest reason we should keep a healthy distance from strangers is precisely due to their unpredictability, to the possibility that they will do us great harm. A properly programmed robot agent wouldn't pose such a threat. There are likely other reasons we tend to keep a certain distance from strangers, however, reasons that robot agents may or may not be able to overcome. On the point of whether we could have loving relationships with robots, Sven Nyholm and Lily Frank argue quite convincingly that "what goes on on the inside matters greatly" [28:223]. These issues do not concern me here. My claim concerns specifically the larger societal moral community for which the membership requirements are quite low (and yet far from fair[16]). For although we may not wish to or be able to accept robots into our intimate moral communities, a properly programmed robot could pass the societal moral membership test with flying colors.

4. Objections

I shall now turn to two possible objections even should one accept my claims (a) about the kind of moral agency our moral responsibility system reflects and fosters and (b) that robots can achieve that standard. The first one is that the moral responsibility responses still have an advantage over any possible robotic programming in that they can address new behaviors that arise in contexts which could not have been foreseen by the programmers of robot agents. Since mistakes in rule/value compliance may be unavoidable among robots (though the most harmful mistakes could presumably be avoided entirely), it is conceivable to implement a relatively primitive reinforced learning system that functions possibly more reliably than the moral responsibility system – whereby mistakes are negatively reinforced and moral actions are positively reinforced (see [29] for an example).

A second possible objection is that the moral responsibility system (i) is not meant to capture or cultivate the moral agency really at work in our moral lives and (ii) is not meant to do nor can it be expected to do the job of moral formation. Regarding (i), I agree. First, human behavior is complex and a definition of moral agency is far from widely agreed upon. Whatever the true picture of how we reason and why we behave the way we do, however, the moral responsibility system defines the requirements for moral agency; it determines our status as members of the moral community. Accordingly, what a human's moral agency actually consists in is irrelevant to our membership; our behavior is what counts. Second, due to the very high demands on any desert attribution, moral responsibility criteria are necessarily very low. We cannot justify punishment unless we have a very high degree of confidence in an agent's guilt. Such confidence is only possible when the standards are sufficiently low. If moral responsibility were

16 These requirements, however minimal, are not made uniformly available or accessible to all due to unequal conditioning practices, disparate values, and dramatically unequal environments.

deemed possible only for an autonomous agent (following [18]), for example, our moral responsibility practices might rarely if ever be justified. Consequently, irrespective of our current or potential capacities for higher degrees of moral agency, our moral responsibility practices both rely on and cultivate the most minimal standards.

To (ii) I respond that although the moral responsibility system may not be intended to function to this end, it does. Whether or not when we blame or punish we are intending to reform someone, someone whom we may even believe to be beyond reform, when we do, we cannot but pair the agent's behavior with our negative reaction. Whether or not the agent is successfully conditioned by our actions and attitudes, we have engaged in an act of conditioning. If that agent is thereby reformed, it is on that limited basis. Take a clear-cut case of behavioral conditioning. If the dog's electric bark collar zaps the dog's neck every time he barks and it hurts, the dog will refrain from barking in the future while he is wearing his bark collar (and the hope is even when he's not). If a child's hand is slapped every time she reaches for a cookie in the cookie jar without first receiving permission, she will likely stop reaching for the cookie without asking first, at least while in view of the slapper. We cannot expect for the child formed in this way to understand why it is important to ask for permission only that it is; "it's important to follow this rule because a source of authority said so and because I will get a hand slap if I don't."

One might respond that we must therefore equip her with the reason – an easy enough task, surely? Not so. Not only must the child be in possession of the reason, she must care about that reason. The rule that children must ask for permission before grabbing and consuming things, which is a rule because of its relationship to health as relates to hygiene and food, may not motivate a child until she understands the reasons for the rule – dirty hands and excessive sugary foods affect health – and perhaps not just in an abstract way, in a direct, personal, possibly physiological way. It may be the case (see [30]) that we are motivated to act correctly (e.g. to stop smoking when we "know" it's unhealthy) when we can experience, even if only in an imagined way, the effects of a particular action. Since it will take time and careful learning methods for the child to form this understanding, rules and the consequences imposed by rule enforcers, along with positive and negative reinforcement, serve to mostly keep children (and adults) from wreaking havoc and causing severe harm. Our moral responsibility and legal responses only come into play when these rules are not being followed. They do not educate in a truly forward-looking way and by relying on them for moral formation, we may even risk cultivating the wrong standards for moral agency; for although we cannot force the child (or the adult) to care about the effects of her actions (which she cannot do if she doesn't fully understand them), we could cause her to care less about them (e.g. the health import of the permission-asking rule) if we direct her attention to the secondary consequences of punishment (e.g. the hand slap).

5. Conclusion

Our moral responsibility system thus, at least partly, blocks the formation of "substantive moral agency" by, *inter alia*, drawing our attention to and prioritizing secondary consequences (e.g. punishment and reward) over and at the cost of a sensitivity to the moral reason in the interest of ensuring behavioral compliance. It narrows and limits the paths we can take in life, should we be socially sensitive or care about a certain kind of social success, and thus also precludes the kind of freedom that might be conceived of as having meaningfully open alternatives. From the society's perspective it

is not relevant why the agent acts, only that the agent performs according to our expectations. On Strawson's view, it's the fulfilling of these expectations that makes community possible. Agents who do – flesh and blood or synthetic – are thus moral community members. Given the primacy of community for morality and the life worth living, prioritizing behavioristic moral agency is understandable; ensuring a minimum of rule observance is vital.

The intuitively objectionable scenario in which a human agent, uncoerced, refrains from wrongdoing (e.g. an act of vandalism) for other than the moral reason (due to an apparent security camera), will nonetheless go uncriticized in practice for we tend only to blame, punish, or resent wrong actions. But such a scenario is not objectionable vis-à-vis robots. Constraints on action are built directly into their system. We want the robot to follow the rule because it is programmed to do so. We care about humans' motives for action; we want to do and to want to do the right thing because it is right and, on some level, we want for that to be a product of a free choice. It's here that substantive moral agency matters. Although it may not matter to our daily, societal-level interactions, it matters to the quality of life we lead, to the quality of our interactions and collaborations, and to the partnerships that we form. Interestingly, the necessary (though perhaps insufficient) condition for those interactions is trust. We must trust that our values will be upheld and this level of trust, I believe, is already attainable by robots. But we must also trust that our values are shared. Can we say that a robot, by virtue of abiding by our values, shares them? Can or could a robot help develop those values or formulate new, progressive goals arising from them?

In conclusion, if we find the behavioristic conception of moral agency wanting, if what goes on on the inside does or should matter, we ought to change our moral responsibility standards. While it would be unfair and unproductive to raise the bar for the purposes of backward-looking accountability, a bar I have argued is already more reliably reached by robots, we should raise it for forward-looking purposes, at least for humans. Concerning membership into the larger moral community, Amanda Sharkey's proposal [8] for the prioritization of safety may be better-headed. We need, first and foremost, safe humans, safe pets, and safe robots. In order for robots to be safe enough in general social settings,[17] they need to follow the rules, which I believe they can do much more reliably than organic agents. Our current means of attempting to render human beings safe seems to come at the cost of substantive agency. Finally, were we to reject the standards of moral agency the moral responsibility practices rely on and foster, we should also retract membership from the significant number (the larger portion, I fear) of human agents in whom society has failed to cultivate a more substantive moral agency. In this respect, many of current society's "card-carrying" human moral agents stand or fall together with the robots.

References

[1] Danaher, J. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. Sci Eng Ethics (2019). https://doi-org.libproxy.helsinki.fi/10.1007/s11948-019-00119-x

17 Not in all settings; after all, most of us would not be safe as more than observers in settings in which we lack sufficient knowledge and skills (e.g. car repair shops, construction sites, science labs, hospital operating theaters, child-rearing, etc.).

[2] Coeckelbergh M. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. AI & Soc. 2009 Sep;24(2):181–9.
[3] Floridi L, Sanders JW. On the morality of artificial agents. Minds and Machines. 2004 Aug;14(3):349–79.
[4] Sullins JP. When is a robot a moral agent. International Review of Information Ethics. 2006;6(12):23–30.
[5] Powers TM. Prospects for a kantian machine. IEEE Intell Syst. 2006 Jul;21(4):46–51.
[6] Arkin RC. Governing lethal behavior in autonomous robots [Internet]. Boca Raton, Fla.: Chapman & Hall/CRC; 2009 [cited 2020 Jun 19]. Available from: http://www.crcnetbase.com/isbn/9781420085945
[7] Arkin RC, Ulam P, Wagner AR. Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc IEEE. 2012 Mar;100(3):571–89.
[8] Sharkey A. Can robots be responsible moral agents? And why should we care? Connection Science. 2017 Jul 3;29(3):210–6.
[9] van Wynsberghe A, Robbins S. Critiquing the reasons for making artificial moral agents. Sci Eng Ethics. 2019 Jun 1;25(3):719–35.
[10] Strawson PF. Social morality and individual ideal. Philosophy. 1961 Jan;36(136):1–17.
[11] Strawson PF. Freedom and resentment and other essays. London ; New York: Routledge; 2008. 235 p.
[12] Wallace RJ. Responsibility and the moral sentiments. Cambridge: Harvard Univ. Press; 1998. 275p.
[13] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Minds & Machines. 2018 Dec;28(4):689–707.
[14] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.
[15] Gunkel DJ. The machine question: critical perspectives on AI, robots, and ethics. Cambridge: MIT Press; 2012. 256 p.
[16] Gunkel DJ. Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol [Internet]. 2017 Jul 19 [cited 2020 Jun 19]; Available from: http://link.springer.com/10.1007/s10676-017-9428-2
[17] Bryson JJ. Robots should be slaves. In: Wilks Y, editor. Natural Language Processing [Internet]. Amsterdam: John Benjamins Publishing Company; 2010 [cited 2020 Jun 20]. p. 63–74. Available from: https://benjamins.com/catalog/nlp.8.11bry
[18] Mele A. Autonomous agents: from self-control to autonomy. Oxford: Oxford Univ. Press; 2001. 271 p.
[19] Hakli R, Mäkelä P. Robots, Autonomy, and Responsibility. In: Seibt J, Nørskov M, Andersen S editors. What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016. Amsterdam, The Netherlands: IOS Press. pp. 145-154.
[20] Irrgang B. Ethical acts in robotics. Ubiquity. 2006 Sep 1;2006(September):2–16.
[21] Frankfurt HG. Freedom of the will and the concept Rice AS, Farquhar-Smith WP, Bridges D, Brooks JW. Canabinoids and pain. In: Dostorovsky JO, Carr DB, Koltzenburg M, editors. Proceedings of the 10th World Congress on Pain; 2002 Aug 17-22; San Diego, CA. Seattle (WA): IASP Press; c2003. p. 437-68 of a person. The Journal of Philosophy. 1971 Jan 14;68(1):5.
[22] Fischer JM, Ravizza M. Responsibility and control: a theory of moral responsibility. Cambridge ; New York: Cambridge University Press; 1998. 277 p. (Cambridge studies in philosophy and law).
[23] Nelkin DK. Making sense of freedom and responsibility. Oxford ; New York: Oxford University Press; 2011. 194 p.
[24] Skinner BF. Beyond freedom and dignity. 1st ed. New York: Knopf; 1971. 225 p.
[25] Piaget J. The moral judgment of the child. 1st pbk. ed. London ; New York: Routledge; 2013. 418 p. (International library of psychology ; Developmental psychology).
[26] Waller BN. Against moral responsibility. Cambridge, Mass: MIT Press; 2011. 352 p.
[27] Himma KE. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics Inf Technol. 2009 Mar;11(1):19–29.
[28] Nyholm SR, Frank LE. From sex robots to love robots: is mutual love with a robot possible? In Danaher J, McArthur N, editors, Robot sex: social and ethical implications. Cambridge: MIT Press. 2017.
[29] Wallach W, Allen C. Moral machines: teaching robots right from wrong. Oxford ; New York: Oxford University Press; 2009. 275 p.
[30] Mischel W. The marshmallow test: mastering self-control. First edition. New York: Little, Brown and Company; 2014. 328 p.